

# **AIR QUALITY PREDICTION USING ARIMA MODEL**

## **A PROJECT REPORT**

*Submitted by*

SAKET SAGAR, Roll No.- 1261900781, Reg. No.-014954

NIKKI PRIYA, Roll No.-12620002063, Reg. No.-201260100220005

SUNDARAM KUMAR, Roll No.– 12619002058, Reg. No.-017108

*Under the Supervision of*

**Prof. (Dr.) DEBABRATA DATTA**

(Professor of Dept. of IT & Joint Director Research & Development and former BARC  
Scientist)

In partial fulfillment for the award of the degree

*Of*

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

**HERITAGE INSTITUTE OF TECHNOLOGY, KOLKATA**

**MAULANA ABUL KALAM AZAD UNIVERSITY OF  
TECHNOLOGY, KOLKATA**



## **CONTENTS**

<b>BONAFIDE CERTIFICATE .....</b>	<b>5</b>
<b>ACKNOWLEDGEMENT .....</b>	<b>6</b>
<b>LIST OF FIGURES .....</b>	<b>7</b>
<b>CHAPTER -1 .....</b>	<b>8</b>
<b>INTRODUCTION... ..</b>	<b>9-11</b>
1.1 MOTIVATION .....	11
1.2 LITERATURE REVIEW .....	12-16
1.3 GAP AREAS .....	16
<b>CHAPTER -2 .....</b>	<b>17</b>
<b>PROBLEM STATEMENT .....</b>	<b>18</b>
2.1 METHODOLOGY .....	18
2.2 FLOW CHART .....	19
<b>CHAPTER – 3 .....</b>	<b>20</b>
<b>PROCESS .....</b>	<b>21</b>
3.1 DATA COLLECTION .....	21
3.2 DATA CLEANSING .....	21
3.3 DATA VISUALIZATION .....	21-22
3.4 TRAINING DATA SET FORMATION .....	24
<b>CHAPTER -4 .....</b>	<b>25</b>
<b>ARIMA MODEL .....</b>	<b>26</b>
4.1 AUGMENTED DICKEY FULLER TEST(ADF TEST) ..	27
4.2 AUGMENTED TEST ON DATA SET .....	27
4.3 IMPLEMENTING AUTO ARIMA FOR P,D,Q .....	28
4.4 IMPLEMENTING AUTO ARIMA ON DATA SET .....	28
<b>CHAPTER -5 .....</b>	<b>29</b>
<b>RESULT .....</b>	<b>30</b>

5.1 PREDICTION USING ARIMA .....	30-31
<b>CHAPTER – 6 .....</b>	<b>32</b>
<b>OBSERVATION.....</b>	<b>33</b>
6.1 THE INFLUENCE OF WEATHER FACTORS ON ARIMA MODEL IN AQI PREDICTION.....	34-35
6.2 HURST EXPONENT .....	36
6.3 FRACTUAL DIMENSION.....	36
<b>CHAPTER – 7 .....</b>	<b>37</b>
<b>CONCLUSIONS.....</b>	<b>38</b>
7.1 ADVANTAGES .....	39
7.2 FUTURE SCOPE.....	39
7.3 REFERENCES.....	40



**HERITAGE INSTITUTE OF TECHNOLOGY**  
**KOLKATA**

*An Autonomous Institution under*

**MAULANA ABUL KALAM AZAD UNIVERSITY OF  
TECHNOLOGY**

**BONAFIDE CERTIFICATE**

Certified that this project report on “**AIR QUALITY PREDICTION USING ARIMA MODEL**” is the Bonafide work of “**SAKET SAGAR, NIKKI PRIYA AND SUNDARAM KUMAR**” who carried out the project under my supervision.

**SIGNATURE**

**Prof. (Dr.) Siuli Roy**

**HEAD OF THE DEPARTMENT**

**Information Technology,**

**Heritage Institute of Technology,**

**Kolkata – 700107.**

**SIGNATURE**

**Prof. (Dr.) Debabrata Datta**

**PROJECT MENTOR**

**Information Technology,**

**Heritage Institute of Technology**

**Kolkata – 700107.**

**SIGNATURE**

**EXTERNAL EXAMINER**

---

## **ACKNOWLEDGEMENT**

We would like to take this opportunity to thank **Prof. (Dr.) Siuli Roy**, Head of the Department of Information Technology for giving me the opportunity to work on this project.

I would also like to thank **Prof. (Dr.) Debabrata Datta (Mentor)** (Professor of Department of IT & **Joint Director Research & Development of HITK and former BARC Scientist**) for constantly supporting us and guiding us throughout the project. Their guidance and words of encouragement motivated us to achieve our goal and impetus to excel.

We thank our faculty members and Laboratory assistants at the Heritage Institute of Technology for paying a pivotal and decisive role during the development of the project. Last but not the least we thank all friends for their cooperation and encouragement that they have bestowed on us.

With thanks to all,

Saket Sagar

Nikki Priya

Sundaram Kumar

---

### List of Figures

<b>Figure No.</b>	<b>Figure Title</b>	<b>Page No.</b>
Figure – 3.3.1	Air Quality Index from Year 2015-2020	Page – 21
Figure – 3.3.2	Monthly Average Emission from the Year January 2015- June 2020	Page – 22
Figure – 3.3.3	Daily Emission from January 2015 – June 2020	Page – 22
Figure – 3.3.4	Monthly Average Emission from the year Jan 2015 – June 2020	Page - 23
Figure – 4.2.1	ADF Test	Page – 27
Figure – 4.4.1	Implementing Auto ARIMA on Dataset	Page – 28
Figure – 5.1.1	Predictions using ARIMA Model	Page – 30
Figure – 5.1.2	AQI Comparisons	Page – 30
Figure – 5.1.3	Predictions	Page – 31
Figure – 6.2.1, 6.2.2	Hurst Exponent	Page – 35
Figure -6.3.1	Fractal Dimension	Page - 36

---

# CHAPTER -1 - INTRODUCTION



## **1.INTRODUCTION**

---

Contamination of the air, in particular in metropolitan areas, is a very well-known problem. The ever-growing population of cities and the increasing level of motorization contribute to the ever-increasing traffic volume, and consequently, the ever-increasing exhaust gases emissions. At the same time, the thickening of city buildings reduces ventilation and increases the porosity of surface, which ends up decreasing the effect of the wind on the evacuation of contamination. The typical sources of air pollution are well-known, but difficult to eliminate, at least completely. Thus, most studies are focused on determining the impact of factors that may modify the concentrations of contaminants in the atmosphere such as transformation, retention or evacuation.

In an attempt to make air quality measurement easier to understand, the ministry of environment and forests launched a National Air Quality Index (AQI). It will put out real time data about level of pollutants in the air and inform people about possible impacts on health.

The air quality index (AQI) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short time period. The purpose of the AQI is to help people know how the local air quality impacts their health. The Environmental Protection Agency (EPA) calculates the AQI for five major air pollutants, for which national air quality standards have been established to safeguard public health.

- 1. Ground-level ozone**
- 2. Particle pollution/particulate matter (PM<sub>2.5</sub>/pm 10)**
- 3. Carbon Monoxide**
- 4. Sulphur dioxide**
- 5. Nitrogen dioxide**

The higher the AQI value, the greater the level of air pollution and the greater the health concerns. The concept of AQI has been widely used in many developed countries for over the last three decades. AQI quickly disseminates air quality information in real-time.

- **How is AQI calculated?**

India follows that the 500-point scale, wherein rating between 0 and 50 is considered good. Rating between 301 to 500 range is deemed hazardous. Every day monitors record concentrations of the major pollutants. These raw measurements are converted into a separate AQI value for each pollutant (ground-level ozone, particle pollution, carbon monoxide, and sulphur dioxide) using standard formulae developed by EPA. The highest of these AQI values are reported as the AQI value for that day.

1. **Air Quality Index Categories:**

**Good (0–50)** - Minimal Impact

2. **Satisfactory (51–100)** - May cause minor breathing difficulties in sensitive people.

3. **Moderately polluted (101–200)** - May cause breathing difficulties in people with lung disease like asthma, and discomfort to people with heart disease, children and older adults.

4. **Poor (201–300)** - May cause breathing difficulties in people on prolonged exposure, and discomfort to people with heart disease.

5. **Very Poor (301–400)** - May cause respiratory illness in people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.

6. **Severe (401-500)** - May cause respiratory issues in healthy people, and serious health issues in people with lung/heart disease. Difficulties may be experienced

even during light physical activity.

## **7. Objectives of Air Quality Index (AQI)**

- Comparing air quality conditions at different locations/cities.
- It also helps in identifying faulty standards and inadequate monitoring programmes.
- AQI helps in analyzing the change in air quality (improvement or degradation).
- AQI informs the public about environmental conditions. It is especially useful for people suffering from illnesses aggravated or caused by air pollution.

### ***1.1 Motivation behind making the project***

When it comes to human health, clean air is the most basic commodity. Today, poor air quality is one of the leading causes of a variety of severe health problems. To estimate the impact on our health, we must be aware of the air quality in our neighborhood, city, and country. The intricate interaction of various components, including chemical reactions, climatic aspects, and emissions from natural and manmade sources, results in air quality. The levels of air pollution in most urban areas have been a source of severe worry. People have the right to know the quality of the air they breathe. However, the data collected by the National Ambient Air Monitoring Network is reported in a format that is difficult to comprehend by the average person, and hence the current air quality information system does not support people's participation in air quality improvement activities.

The main goal of this project is to investigate the state and quality of the air by measuring the Air Quality Index (AQI) and comparing the measured values to standard values in order to create environmental impact. Therefore, building a forecasting system for predicting the air quality based on the levels of concentration of individual pollutants and various meteorological parameters will be useful for the population's health.

---

## *1.2.Literature Survey*

### **1.2.1.Prediction of Air Quality Index and Forecasting Ambient Air Pollutants using Machine Learning Algorithms [1]**

---

#### **Author-**

- Sonali. K. Powar Assistant Professor, Modern College, Ganeshkhind, Pune (M.S.) INDIA
- Dr. H. T. Dinde I/C Principal, Karmaveer Bhaurao Patil College, Urun – Islampur (M.S.) INDIA
- Radhika M.Patil B.Sc.CS (Entire) department, Vivekanand College, Kolhapur (Autonomous) INDIA.

**Published on** – 8 August- 2020

The purpose of this literature review paper is to know in detail about the Air Quality Index (AQI) as AQI tells whether the air around us is polluted or not. It is important to know about AQI because unless and until the people know the worst impacts or hazards of air pollution they will not become that much aware about the air pollution and try to reduce it. As per this review most of the researchers worked on AQI and pollutants concentration level forecasting that will give the actual idea about AQI. Artificial Neural Network (ANN), Linear and Logistic Regression are the choices of many researchers for the prediction of AQI and air pollutants concentration.

**Future work** - The future scope may include consideration of all parameters that is meteorological parameters, air pollutants while predicting AQI or forecasting

the future concentration level of different pollutants.

### **1.2.2. Multi-Model Federated Learning: An Advanced Approach for Air Quality Index Forecasting [2]**

---

**Aurthor** - Lê Đồng(University of Economics Ho Chi Minh City), Anh-Khoa Tran, Minh Dao, (National Institute of Information and Communications Technology), Kieu-Chinh Nguyen-Ly

**Published on** – November 2022

In this model, Statistical models have played a massive role in predicting air quality, especially statistical models with the application of AI techniques. ANNs have been shown to outperform traditional statistical models that do not use AI. Unfortunately, the biggest drawback of this type is that it often falls into local optimal stages. To overcome this draw- back, many studies have been used, such as DNN hybrid and ensemble. The AQI forecast should assess the implications of air pollution on several sectors, such as health, agriculture, land transportation, aviation, and energy, among others, and issue warnings and recommendations based on various thresholds, risks, and cost functions. Some works concentrate on RNN and the Spatial-Temporal Network can deal with the complicated non-linear spatial and temporal correlations.

#### **Conclusion –**

We have gathered several AQI prediction studies as well as FL research in this survey. We discovered through research and synthesis that, prior to 2020, the majority of AQI forecasting mechanisms relied on single ML techniques such as statistics, ANN, DNN, hybrid, and ensemble. Although commonly used, ANN and DNN were prone to failure in local optimum conditions. The anticipated outcomes could be enhanced by using some

hybrid and ensemble approaches.

Since 2000, various works have integrated FL into AQI processing. We have identified the three different FL architectures: centralized FL, decentralized FL, and hierarchical FL as well as their corresponding algorithms. The system mostly used centralized FL, and the typical process model was DNNs. There has been a trend in the use of UAVs in this field. Additionally, we mentioned several benchmarks, challenges and solutions, and domains of FL. Reviews demonstrate that multi-model FL improves the accuracy of other domains, thus it could be applied to enhance AQI predictions. Multi-model FL is a study area that needs to be focused on in future works. DNN should be incorporated into client models, since it is more responsive than conventional ML architectures. Additionally, DNN should be designed to adopt both partial and temporal data.

More importantly, to have a common baseline to assess new solutions, the scientific community should also be interested in developing a dedicated open dataset on AQI.

### **1.2.3. Air pollution prediction with machine learning: a case study of Indian cities [3]**

**Author** - Kaushal Kumar (Guru Nanak Dev University), Dr. B. P. Pande (LSM GPGC Pithoragarh)

**Published on** - May 2022

In this model, The air pollution dataset is splitted into training(75%) and testing(25%) subsets before evaluating ML models. To examine the seasonality of the data thoroughly, Box plot visualization are employed. Box plot categories data into different periods by grouping the entire information in years and months.

Many ML models ignore this imbalanced datasets problem which may lead to poor classification and prediction performances. To overcome this data imbalance problem, the SMOTE (Synthesis Minority Oversampling Technique) has been applied. To implement SMOTE for class imbalance, they have used an imbalanced-learn python library in the SMOTE class. Now they have used 5 popular ML Models, KNN, Gaussian

Naïve Bayes(GNB), SVM, RF and XGBoost have been employed to predict the AQI level with SMOTE and without SMOTE resampling technique.

**Limitations** – Prediction of air quality is a challenging task because of the dynamic environment, unpredictability and variability in space and time of pollutants. The present research endeavours to contribute to the literature by addressing air quality analysis and prediction for India which might have not been properly studied. This work can be extended by employing deep learning techniques for AQI prediction.

#### **1.2.4. Air quality prediction models based on meteorological factors and real-time data of industrial waste gas [4]**

**Author** \_ Ying Liu, Peiyu Wang, Yong Li, Lixia Wen & Xiaochao Deng

**Published on** – 03 June 2022

In this model, a random forest model is used to construct an air quality prediction model in Zhangdian District based on the real-time dynamic emission effect of industrial waste gas on air quality in the region.

Using this model, the daily emission limit of industrial pollution can be determined according to the weather forecast inversion, and the air pollution risk caused by unfavorable meteorological factors can be effectively avoided by adjusting the production capacity of the internal production process of the enterprise. This research actively responds to the “Fourteenth Five-Year Plan for National Economic and Social Development of Zhangdian District and the Outline of Vision 2035”: by promoting the implementation of typical production scenarios, empowering actions, focusing on digital industrial applications, using cloud computing, big data and other new-generation information technologies, and guidelines for building a new industrialized strong city in the country. It provides a new idea for Zhangdian District’s “14th Five-Year Plan” to achieve an average annual growth rate of regional GDP of more than 7% and the

harmonious development of industry and environment.

**Conclusion** - By comparing the random forest algorithm with other machine learning algorithms, we can verify the applicability of the random forest algorithm for air quality prediction in Zhangdian District. In this paper, four kinds of machine learning algorithms were used to predict AQI, and their results were compared to ascertain the most appropriate machine learning algorithm. The RMSE, MAE, and  $R^2$  measures were used to evaluate the prediction accuracy of the four machine learning algorithms.

For these algorithms, lower RMSE and MAE values indicate higher prediction accuracy, while the closer the  $R^2$  value is to 1, the more accurate the prediction is. The results confirm that the prediction accuracy of the random forest model is better than the other three machine learning models, indicating that the random forest model is the most suitable algorithm for the AQI prediction model of Zhangdian District.

### ***1.3 GAP AREAS***

- The accuracy of the model was found to be unsatisfactory.
- The main reason for the low accuracy is the absence of weather condition data in the model.
- The project did not incorporate weather variables such as temperature, humidity, wind speed and precipitation, which have significant impact on air quality.
- The exclusion of weather conditions limited the model's ability to capture the complex dynamics and interactions influencing air quality.
- The model's inability to account for weather-related factors resulted in inaccurate predictions and hindered its effectiveness as a reliable air quality forecasting tool.
- To enhance the accuracy of future air quality prediction model, it is crucial to integrate weather data to better capture the comprehensive factors affecting air quality.



## CHAPTER 2 – PROBLEM STATEMENT AND METHODOLOGY

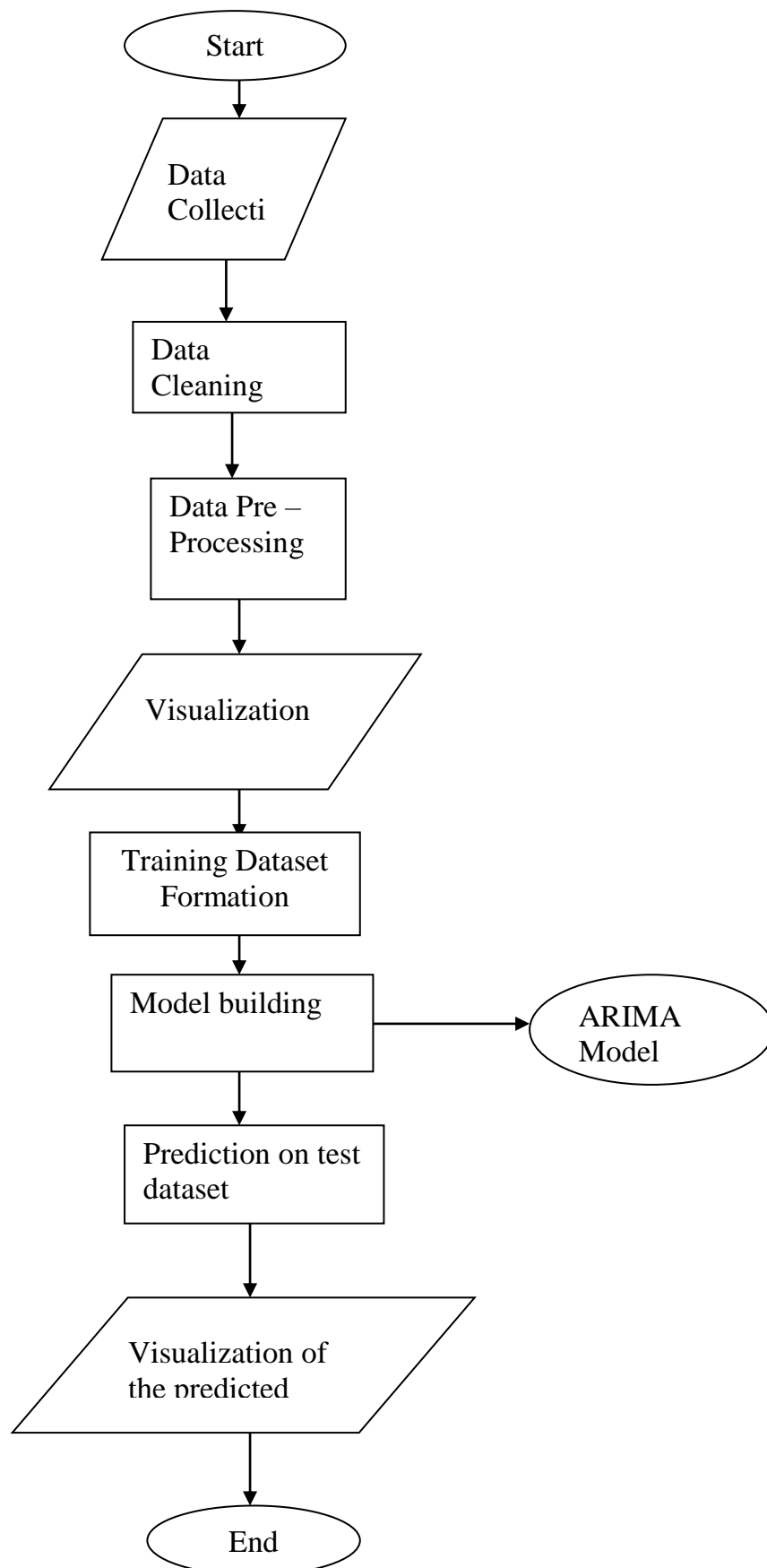
## **2. Problem Statement**

In this project our main aim is to develop an efficient approach for forecasting the air quality index of Bangalore using meteorological parameters and concentration of major air pollutants using various Machine Learning Algorithms and Deep Neural Networks like LSTM.

### **2.1. Methodology**

1. Data Collection
2. Data Cleaning
3. Data Pre-processing
4. Visualization
5. Training Dataset Formation
6. Model Building
7. Prediction on testing dataset
8. Visualization of predicted model

## 2.2. Flow Chart



## CHAPTER 3 – DATA PROCESSING

### 3.1. Data Collection

Concentration of various Air Pollutants like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3 were used. Data was collected from Kaggle.

### 3.2. Data Cleansing

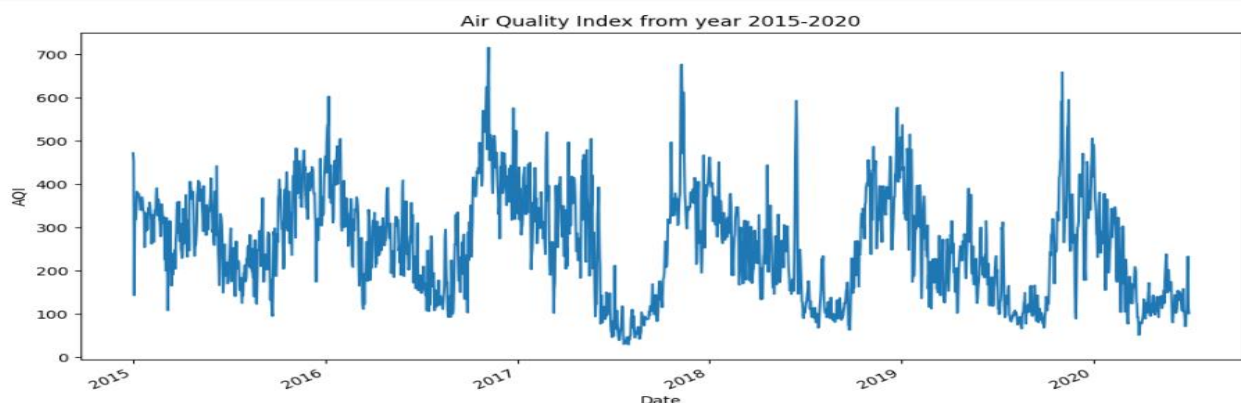
Data cleansing entails identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

From the data, various features which are not useful in AQI prediction like 'NH3', 'NO', 'Benzene', 'Toluene', 'Xylene', 'AQI Bucket' are removed. Finally, the independent features used are PM2.5, PM10, NO2, NOx, CO, SO2 and O3.

Rows with missing AQI values were dropped whereas the missing values in independent features were interpolated.

### 3.3. Data Visualization

The data obtained after cleansing was then visualized to check whether the data is proper or not.



*Figure-3.3.1*  
*Air Quality Index from year 2015 - 2020*

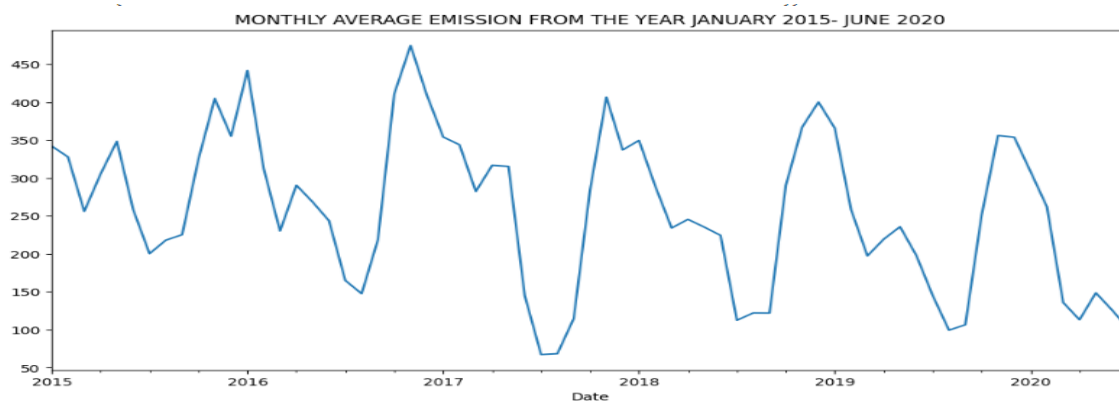


Figure-3.3.2

Monthly Average Emission From the year January 2015 – June 2020

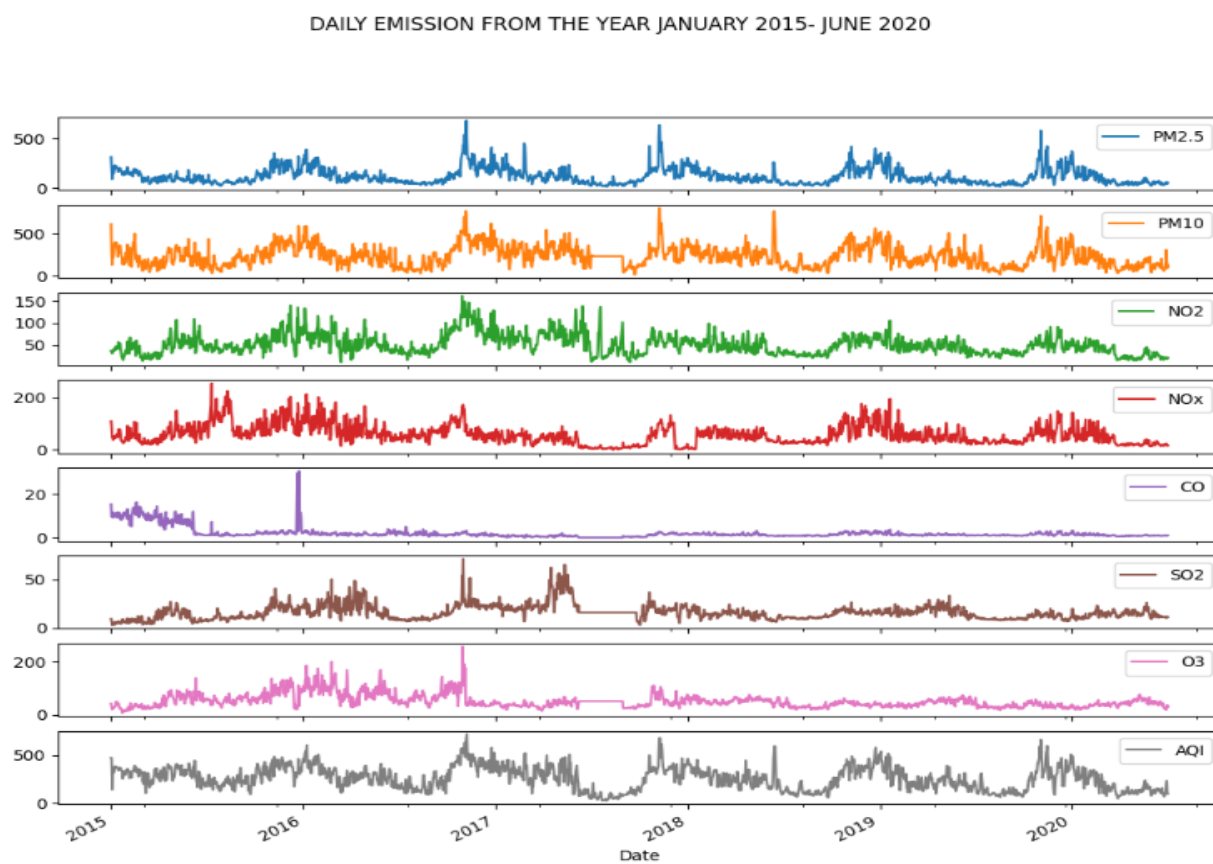
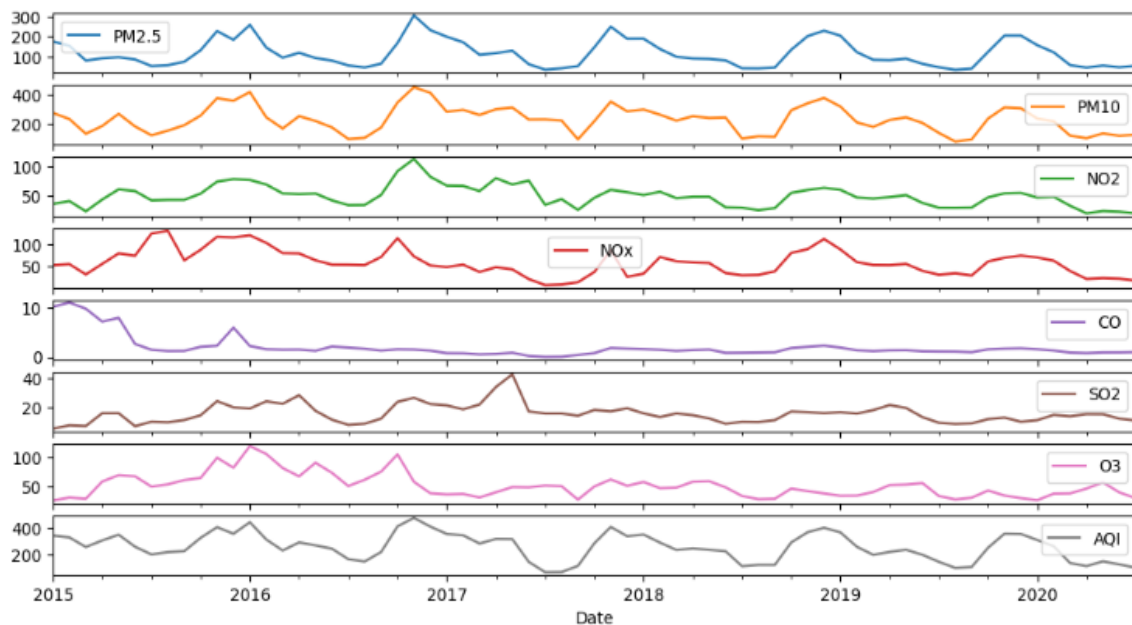


Figure- 3.3.3

Daily Emission From the year January 2015 – June 2020



MONTHLY AVERAGE EMISSION FROM THE YEAR JANUARY 2015- JUNE 2020



*Figure-3.3.4*  
*Monthly Average Emission From the year January 2015 – June 2020*

### **3.4. Data Preprocessing**

The first and most important criterion to ensure the effective construction of forecasting models is data quality and representativity. The capacity of a machine learning algorithm to generalize is often influenced by the data preparation phase. Missing data imputation, eliminating or changing outlier observations, data transformation (typically normalization and standardization), and feature engineering are all examples of data preparation. While the first two phases are important for obtaining more precise and full data sets, the third step is often used to obtain data that is more consistently distributed and to reduce data variability. Finally, the fourth phase is utilized to generate a new dataset that is often smaller and more informative. Feature extraction and feature selection are usually included in the final stage.

### **3.5 Training Dataset Formation**

Data is collected for the years 2016-2020, out of which data from 2016-2019 will be used for training the model and the collected AQI values for 2020 will be compared with the values forecasted using the ARIMA model and various other machine learning models as well.



## CHAPTER 4 – ARIMA MODEL

#### **4. ARIMA Model**

The **ARIMA (AutoRegressive Integrated Moving Average)** [5] model is a popular time series forecasting method that combines **autoregressive (AR)**, **differencing (I)**, and **moving average (MA)** components. It is widely used for analyzing and predicting data with temporal dependencies. The ARIMA model assumes that a time series can be represented as a combination of its own past values (autoregressive component), the difference between current and past values (differencing component), and a moving average of past forecast errors (moving average component). These components capture the trend, seasonality, and random fluctuations in the data.

The ARIMA [6] model is characterized by three parameters: **p, d, and q**. The parameter **p** represents the order of the autoregressive component, **d** represents the degree of differencing, and **q** represents the order of the moving average component. By selecting appropriate values for these parameters, the ARIMA model can capture the underlying patterns and dynamics in the time series.

The ARIMA model is often used for tasks such as time series forecasting, trend analysis, and anomaly detection. It is implemented in various software libraries and packages, such as stats models in Python, and can be applied to a wide range of domains, including finance, economics, meteorology, and more.

It's important to note that while the ARIMA model is a powerful tool for time series analysis, its performance depends on the quality and stationarity of the data, appropriate parameter selection, and the absence of significant outliers or structural breaks in the series.

In summary, the ARIMA model is a flexible and widely used approach for time series analysis and forecasting, combining autoregressive, differencing, and moving average.

## 4.1 Augmented Dickey Fuller Test (ADF Test)

The Augmented Dickey-Fuller (ADF) [10] test is a statistical test used to determine the stationarity of a time series. Stationarity is an important assumption in many time series analysis techniques, as it ensures that the statistical properties of the series remain constant over time.

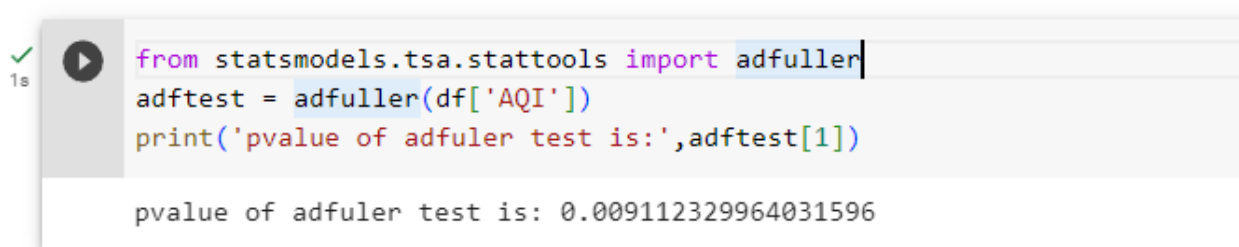
The ADF test is an extension of the Dickey-Fuller test, which tests the null hypothesis that a unit root is present in a time series (indicating non-stationarity). The ADF test goes beyond the basic Dickey-Fuller test by including additional lagged differences of the series to account for potential autocorrelation.

The ADF test evaluates the significance of the coefficient of the lagged differences in a regression model that includes both lagged differences and lagged levels of the time series. The test provides a test statistic and p-value, which can be compared against critical values to make a decision on the stationarity of the series.

If the p-value is less than a predetermined significance level (e.g., 0.05), the null hypothesis of the presence of a unit root (non-stationarity) is rejected, indicating that the series is stationary. On the other hand, if the p-value is greater than the significance level, the null hypothesis cannot be rejected, suggesting that the series is non-stationary.

## 4.2 ADF Test on Dataset

### Augmented Dickey Fuller Test (ADF Test)

A screenshot of a Jupyter Notebook cell. On the left, there is a green checkmark and a play button icon. The cell contains three lines of Python code: `from statsmodels.tsa.stattools import adfuller`, `adftest = adfuller(df['AQI'])`, and `print('pvalue of adfuler test is:',adftest[1])`. Below the code, the output is displayed: `pvalue of adfuler test is: 0.009112329964031596`.

```
from statsmodels.tsa.stattools import adfuller
adftest = adfuller(df['AQI'])
print('pvalue of adfuler test is:',adftest[1])

pvalue of adfuler test is: 0.009112329964031596
```

*Figure-4.2.1*  
*ADF Test on Dataset*

**From the above p-value (0.0091123...), we concluded that our time-series is stationary.**

### 4.3 Implementing Auto ARIMA to find values of p,d,q [6,7]

Auto ARIMA, short for Automated AutoRegressive Integrated Moving Average, is an algorithmic approach for automatically selecting the optimal parameters for an ARIMA model. ARIMA models are widely used in time series analysis and forecasting, but selecting the appropriate values for the model's order parameters (p, d, q) can be a challenging and time-consuming task.

Auto ARIMA automates the process of parameter selection by systematically evaluating different combinations of parameters and selecting the model that yields the best performance based on a given evaluation criterion (e.g., AIC, BIC). It utilizes an iterative algorithm that considers various combinations of differencing, autoregressive, and moving average components to find the most suitable model for the data.

### 4.4 Implementing Auto Arima on Dataset

```
▶ Performing stepwise search to minimize aic
▶ ARIMA(0,0,0)(1,0,1)[12] intercept : AIC=751.256, Time=0.70 sec
▶ ARIMA(0,0,0)(0,0,0)[12] intercept : AIC=811.611, Time=0.02 sec
▶ ARIMA(1,0,0)(1,0,0)[12] intercept : AIC=721.831, Time=0.31 sec
▶ ARIMA(0,0,1)(0,0,1)[12] intercept : AIC=749.203, Time=0.52 sec
▶ ARIMA(0,0,0)(0,0,0)[12] : AIC=945.014, Time=0.04 sec
▶ ARIMA(1,0,0)(0,0,0)[12] intercept : AIC=764.235, Time=0.12 sec
▶ ARIMA(1,0,0)(2,0,0)[12] intercept : AIC=inf, Time=1.98 sec
▶ ARIMA(1,0,0)(1,0,1)[12] intercept : AIC=714.121, Time=1.20 sec
▶ ARIMA(1,0,0)(0,0,1)[12] intercept : AIC=743.571, Time=0.27 sec
▶ ARIMA(1,0,0)(2,0,1)[12] intercept : AIC=inf, Time=1.29 sec
▶ ARIMA(1,0,0)(1,0,2)[12] intercept : AIC=inf, Time=1.10 sec
▶ ARIMA(1,0,0)(0,0,2)[12] intercept : AIC=inf, Time=0.60 sec
▶ ARIMA(1,0,0)(2,0,2)[12] intercept : AIC=inf, Time=1.42 sec
▶ ARIMA(2,0,0)(1,0,1)[12] intercept : AIC=716.497, Time=1.56 sec
▶ ARIMA(1,0,1)(1,0,1)[12] intercept : AIC=716.541, Time=1.46 sec
▶ ARIMA(0,0,1)(1,0,1)[12] intercept : AIC=730.259, Time=1.37 sec
▶ ARIMA(2,0,1)(1,0,1)[12] intercept : AIC=inf, Time=2.79 sec
▶ ARIMA(1,0,0)(1,0,1)[12] : AIC=718.456, Time=1.38 sec

Best model: ARIMA(1,0,0)(1,0,1)[12] intercept
Total fit time: 18.207 seconds
```

Figure-4.4.1

Implementing Auto Arima on dataset

**From the above, we can observe the best values for p, q and q (1, 0, 0).**

## CHAPTER 5 - RESULT

## 5. Results

### 5.1 Predictions using ARIMA Model

The AQI values for the testing dataset were predicted with RMS Error of 72.6493297.

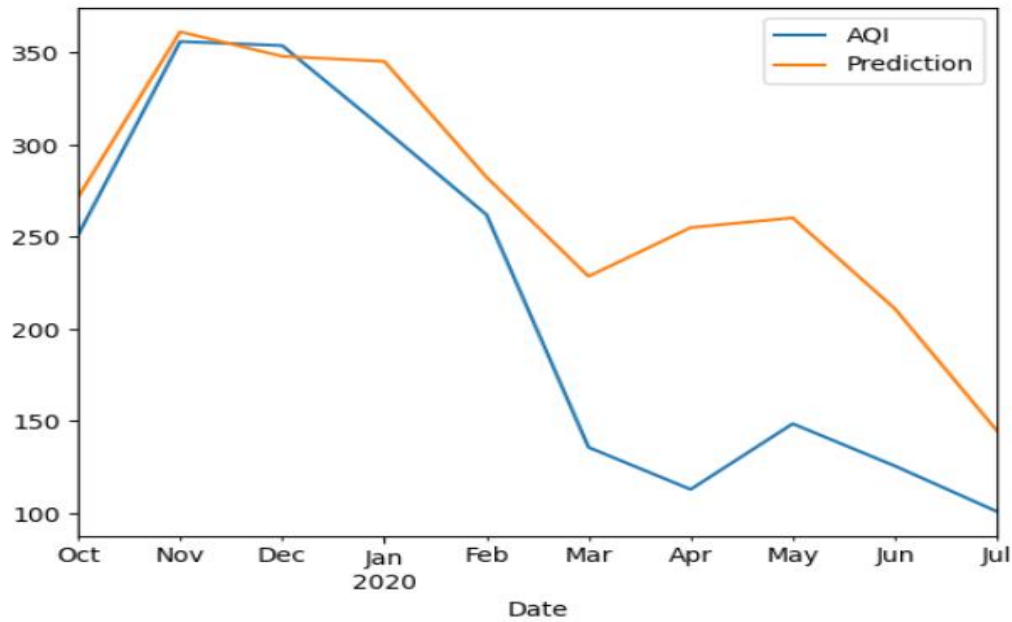


Figure-5.1.1

Prediction 1.1

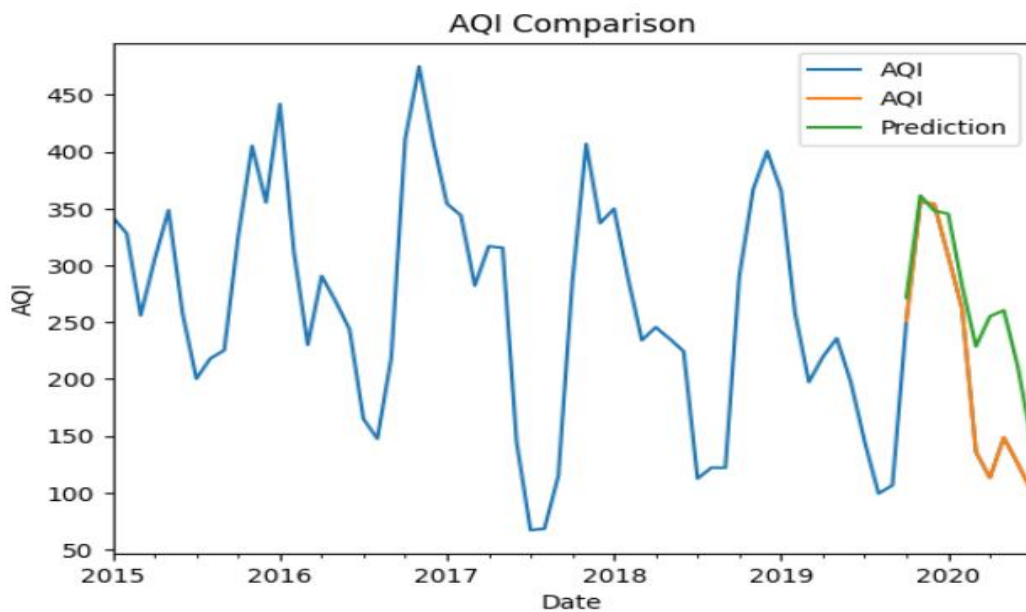


Figure-5.1.2

Prediction 1.2

Prediction 	
Date	
2019-10-31	271.565071
2019-11-30	361.411344
2019-12-31	348.075630
2020-01-31	345.348406
2020-02-29	282.528471
2020-03-31	228.639564
2020-04-30	255.107754
2020-05-31	260.417075
2020-06-30	211.140948
2020-07-31	144.851757

*Figure-5.1.3*  
*Prediction 1.3*

## CHAPTER 6 - OBSERVATIONS



## **6. Observations**

From above, we observed that RMSE is 72.6493297 which is not a good predictions.

### **6.1 The Influence of Weather Factors on ARIMA Model Performance in AQI Prediction**

The ARIMA model is a commonly used technique for time series analysis and forecasting, including air quality index (AQI) prediction. However, it is important to acknowledge that the accuracy and performance of the ARIMA model can be influenced by various factors, including weather conditions. This note aims to highlight the potential impact of weather factors on the effectiveness of the ARIMA model in predicting AQI values.

**Weather Factors and AQI:** Air quality is significantly affected by weather conditions, such as temperature, humidity, wind speed, and atmospheric pressure. These weather factors can influence the dispersion, formation, and transport of air pollutants, resulting in fluctuations in the AQI. Changes in weather patterns can lead to variations in pollutant concentrations and atmospheric dynamics, which may pose challenges to accurately modeling and predicting AQI using the ARIMA model alone.

**Limitations of ARIMA Model in Capturing Weather Effects:** The ARIMA model assumes that the underlying time series is stationary and exhibits a linear relationship between past and future values. While it can capture trends, seasonality, and autocorrelation, it may not adequately capture the complex and non-linear relationship between weather factors and AQI variations. The ARIMA model's inability to directly incorporate weather variables as inputs may limit its predictive power, especially in cases where weather plays a significant role in determining air quality.

**Importance of Weather-Adjusted Models:** To improve AQI prediction accuracy, it is often necessary to consider weather factors explicitly. Weather-adjusted models, such as ARIMA models with weather covariates or more advanced machine learning techniques that incorporate weather data, can be more effective in capturing the influence of weather

on AQI. By including weather variables as additional predictors, these models can better account for the impact of temperature, humidity, wind, and other weather-related parameters on air quality dynamics.

While the ARIMA model is a valuable tool for time series analysis and forecasting, it may not capture the full complexity of the relationship between weather factors and AQI variations. The influence of weather on air quality highlights the need for advanced modeling techniques that explicitly consider weather data and their interactions with pollutant concentrations. By integrating weather-adjusted models and leveraging comprehensive datasets, we can enhance the accuracy of AQI predictions and gain a better understanding of the complex interdependencies between weather and air quality.

## **6.2 Hurst Exponent (H)**

The Hurst exponent [11] is a mathematical measure used to quantify the long-term dependence or persistence of a time series data. It provides valuable insights into the underlying characteristics and behavior of a dataset, particularly in the context of self-similarity and fractal properties.

The Hurst exponent, often denoted as "H," is named after the British hydrologist Harold Edwin Hurst, who introduced it in the 1950s. It is derived from the concept of fractional Brownian motion and is calculated using various methods, such as the R/S analysis or detrended fluctuation analysis (DFA). The Hurst exponent is a value between 0 and 1, with specific interpretations based on its magnitude:

1. **H = 0.5:** A Hurst exponent of 0.5 indicates a random or uncorrelated time series. The data points are independent, and there is no long-term dependence or memory in the series.
2. **H > 0.5:** A Hurst exponent greater than 0.5 suggests positive long-term dependence or persistence. It implies that the time series exhibits a trend or memory, where past values have a significant influence on future values. The larger the value of H (> 0.5), the stronger the persistence.
3. **H < 0.5:** A Hurst exponent less than 0.5 indicates negative long-term dependence

or anti-persistence. It implies that the time series tends to revert to the mean or exhibit mean-reverting behavior. Past values have a weaker influence on future values.

```
Hurst Exponent for the Whole Dataset: 0.7367350885479488
```

*Figure-6.2.1*  
*Hurst Exponent*

```
Hurst Exponent for Year 2015: 0.6701483492968768  
Hurst Exponent for Year 2016: 0.5111576894476765  
Hurst Exponent for Year 2017: 0.45926611124045313  
Hurst Exponent for Year 2018: 0.6735248345710598  
Hurst Exponent for Year 2019: 0.6271253887743313  
Hurst Exponent for Year 2020: 0.6598393946250288
```

*Figure-6.2.2*  
*Hurst Exponent*

**So, from the above data we can conclude that Hurst exponent is greater than 0.5(mostly) and It implies that the time series exhibits a trend or memory, where past values have a significant influence on future values**

### 6.3 Fractal Dimension

Fractal dimension (D) [12] is a measure of the complexity or intricacy of a fractal pattern or object. It quantifies how the detail or structure of a fractal changes as the scale of observation or measurement changes. In short, fractal dimension represents how the object fills or occupies space, exhibiting self-similarity across different scales.

```
➡ Fractal Dimension (D) for Year 2015: 1.329851650703123
   Fractal Dimension (D) for Year 2016: 1.4888423105523234
   Fractal Dimension (D) for Year 2017: 1.5407338887595468
   Fractal Dimension (D) for Year 2018: 1.3264751654289402
   Fractal Dimension (D) for Year 2019: 1.3728746112256687
   Fractal Dimension (D) for Year 2020: 1.340160605374971
```

*Figure-6.3.1*  
*Fractal Dimension*

**In analyzing the data, it has become evident that the underlying patterns exhibit both multifractal and chaotic characteristics. The presence of multifractality implies that the data contains multiple scales of variation, with different regions exhibiting distinct degrees of irregularity and complexity.**

## CHAPTER 7 – CONCLUSION AND REFERENCES

## **9. Conclusion**

The air quality index (AQI) or air pollution index (API) is a standard method of informing the public about the severity of air pollution. Various researchers/environmental agencies have created a number of ways for determining AQI or API in the past, but there is no globally approved method that is adequate for all scenarios. In computing the AQI or API, different methods utilize different aggregation functions and take into account different types and amounts of contaminants.

When dangerous or excessive quantities of specific chemicals such as gases, particles, and biological molecules are introduced into the atmosphere, it is referred to as air pollution. Excessive emissions have apparent repercussions, such as disease and mortality among populations and other living species, as well as crop damage. Because of the dynamic nature, volatility, and great unpredictability in location and time of pollutants and particles, predicting air quality is a difficult undertaking. Simultaneously, due to the recognized significant repercussions of air pollution on humans and the environment, the ability to model, predict, and monitor air quality is becoming increasingly vital, particularly in metropolitan areas.

## **9.1 Advantages**

The ARIMA model has three main parameters -  $p$  – this is called the auto regressive lags which we get from the auto regressive component of the model. We can obtain this parameter using the PACF (partial autocorrelation function) graph.  $d$  – this is called order of differentiation which is the order of differentiation which is required to convert non stationary data to stationary data.  $q$  – this is called the moving average parameter which we get from the moving average component of the model, we can obtain this parameter using ACF (auto correlation function) graph.

On the basis of the train data, we can forecast the future values using ARIMA model. Using the parameter values and the accuracy metrics we can forecast for a whole year and use this forecasting to understand how this can impact the environment in the upcoming years.

Final prediction for forecast and original values is done using ARIMA model with the predicted values. To get more accuracy and exact forecasting of the data, the accuracy metrics of MAE and RMSE would have to be optimized using differentiation. We can further optimize the forecast by increasing the differentiation and accuracy metrics.

## **9.2 Future Scope**

The prediction model can be improved by strengthening the methods to forecast the concentration of air quality factors, majorly for  $O_3$ , as  $O_3$  does not come from direct sources but due to multiple sources of emission and their reaction to each other. There are multiple time series models which can be used for this. The time series data can be collected for two or three years, or more than that, and we can work on that data in order to make more accurate predictions.

### 9.3 References

1. [https://www.researchgate.net/publication/344581083\\_A\\_Literature\\_Review\\_on\\_Prediction\\_of\\_Air\\_Quality\\_Index\\_and\\_Forecasting\\_Ambient\\_Air\\_Pollutants\\_using\\_Machine\\_Learning\\_Algorithms](https://www.researchgate.net/publication/344581083_A_Literature_Review_on_Prediction_of_Air_Quality_Index_and_Forecasting_Ambient_Air_Pollutants_using_Machine_Learning_Algorithms)
2. [https://www.researchgate.net/publication/365491003\\_Insights\\_into\\_Multi-Model\\_Federated\\_Learning\\_An\\_Advanced\\_Approach\\_for\\_Air\\_Quality\\_Index\\_Forecasting](https://www.researchgate.net/publication/365491003_Insights_into_Multi-Model_Federated_Learning_An_Advanced_Approach_for_Air_Quality_Index_Forecasting)
3. [https://www.researchgate.net/publication/360617294\\_Air\\_pollution\\_prediction\\_with\\_machine\\_learning\\_a\\_case\\_study\\_of\\_Indian\\_cities](https://www.researchgate.net/publication/360617294_Air_pollution_prediction_with_machine_learning_a_case_study_of_Indian_cities)
4. <https://www.nature.com/articles/s41598-022-13579-2>
5. <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp>
6. <https://www.capitalone.com/tech/machine-learning/understanding-arima-models/>
7. <https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/#:~:text=Draw%20a%20partial%20autocorrelation%20graph,to%20the%20ACF%20is%20q.>
8. [https://www.researchgate.net/post/How\\_to\\_determine\\_p\\_d\\_q\\_values\\_for\\_ARIMA\\_model](https://www.researchgate.net/post/How_to_determine_p_d_q_values_for_ARIMA_model)
9. [https://www.researchgate.net/publication/337704392\\_Forecasting\\_Air\\_Quality\\_of\\_Delhi\\_Using\\_ARIMA\\_Model](https://www.researchgate.net/publication/337704392_Forecasting_Air_Quality_of_Delhi_Using_ARIMA_Model)
10. <https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>
11. <https://pubsonline.informs.org/doi/10.1287/LYTX.2012.04.05/full/>
12. <https://statusneo.com/fractal-dimensions-for-feature-extraction-in-time-series/>



